Sparse Partial Least Squares Discriminant Analysis (PLSDA) for Classification of Microorganisms Using FTIR Spectroscopy

V. Tafintseva[1], V. Shapaval[1,2], A. Kohler[1]

[1]Dept. Mathematical Sciences and Technology, Norwegian University of Life Sciences
Drøbakveien 31, 1430 Ås, Norway
[2]Nofima AS, Osloveien 1, 1430 Ås, Norway

FTIR spectroscopy in combination with multivariate analysis has been proven a powerful tool for the identification of microorganisms. It is a sensitive biophysical technique, which performs microbial identification at different levels such as genus, species and in some cases even strains. For building multivariate identification schemes, regression trees built on artificial networks have been frequently used [1]. Partial Least Squares Regression (PLSR) methods are multivariate methods, which are widely used for classification or discrimination problems, but which have been only rarely used for building identification trees for FTIR spectroscopy data [2]. PLSR is based on latent variables, which enable biochemical interpretation on the basis of regression coefficients or PLS components, a prerequisite which they do not share with classification schemes based on Artificial Neural Networks (ANN). Regression coefficients and PLSR components provide valuable information on the principal differences among classes. In order to enhance interpretation different variable selection techniques have been introduced for PLSR. Recently, the use of sparse PLSR has gained interest both in the field of genetics and metabolomics and been applied to FTIR spectroscopy data [3, 4]. By sparse PLSR, noise can successfully be suppressed and important chemical bands can be highlighted.

The aim of this study was to investigate the use of sparse PLSR for classification problems. As a test data set, spectra obtained from fifty-nine food spoilage mould strains were used[2, 5]. A regression tree based on a phylogenetic tree consisting of 5 levels (Division, Class, Genus, Sub-genus and Species) and 13 different species in total was established. In each node, sparse PLS discriminant analysis was used. The model achieved high classification results (classification rate 85%) and even more importantly sparse and highly interpretable regression coefficients.

References

[1] T. Udelhoven, D. Naumann, J. Schmitt, *Applied Spectroscopy* 54(10), 1471-1479 (2000).

[2] K.H. Liland, A. Kohler, V. Shapaval, *Chemometrics and Intelligent Laboratory Systems* 138, 41-47 (2014).

[3] I. Karaman, N.P. Nørskov, C.C. Yde, M.S. Hedemann, K.E.B. Knudsen, A.Kohler, *Metabolomics* 11, 367-379 (2015).

[4] I. Karaman, E.M. Qannari, H. Martens, M.S. Hedemann, K.E.B. Knudsen, A. Kohler, *Chemometrics and Intelligent Laboratory Systems* 122, 65-77 (2013).

[5] V. Shapaval, J. Schmitt, T. Møretrø, H.P. Suso, I. Skaar, A.W. Åsli, D. Lillehaug, A. Kohler, *Journal of Applied Microbiology* 114(3), 788-796 (2013).